Progress in evidence based reproductive surgery

J. Bosteels¹,², S. Weyers³, C. Siristatidis⁴, S. Bhattacharya⁵, T. D’Hooghe⁶

¹ Department of Obstetrics and Gynaecology, Imeldahospitaal, Imeldalaan 9, 2820 Bonheiden, Belgium.
² CEBAM, Centre for evidence-based medicine, the Belgian branch of the Cochrane Collaboration, Capucijnenvoer 33, blok J, 3000 Leuven, Belgium.
³ Universitair Ziekenhuis Leuven, University Hospital Gent, De Pintelaan 185, 9000 Gent, Belgium.
⁴ Assisted Reproduction Unit, 3rd Department of Obstetrics & Gynecology, “Attikon” University Hospital, Athens, Greece.
⁵ Aberdeen Maternity Hospital, School of Medicine and Dentistry, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, United Kingdom.
⁶ Department of Obstetrics and Gynaecology, University Hospital Gasthuisberg, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

Correspondence at: Dr. Jan Bosteels, Department of Obstetrics and Gynaecology, Imeldahospitaal, Imeldalaan 9, 2820 Bonheiden, Belgium.
E-mail: jan.bosteels@med.kuleuven.be

Abstract

The Consolidated Standards of Reporting Trials (CONSORT) was introduced in 1996 to improve the methodological quality of published reports of randomised controlled trials. By doing a systematic review of randomised controlled trials on reproductive surgery, our group can demonstrate that the overall quality of the published reports of randomised studies on reproductive surgical interventions has improved after CONSORT. Nevertheless, some problems still remain. By discussing the benefits and pitfalls of randomised trials in reproductive surgery, our opinion paper aims to stimulate the reader’s further interest in evidence-based practice in reproductive surgery.

Key words: Evidence-based medicine, randomised controlled trials, reproductive surgery.

Introduction

Traditionally, a wooden spoon was given every year to the student with the lowest score at the comprehensive mathematics examination at St Johns College in Cambridge University. It was awarded for the last time in 1909. Its possession implied that its owner was actually better equipped to be a cook than a scholar.

In 1979 Archie Cochrane awarded a wooden spoon to Obstetrics and Gynaecology because the uptake of designing randomised controlled trials (RCTs) in this discipline was almost non-existent. Some time before, he had criticized the medical profession by writing that “we have not organised a critical summary, by specialty or subspecialty, updated periodically, of all relevant randomised controlled trials” (Cochrane, 1979). Initially, Cochrane’s challenge was taken up in perinatal medicine. In the field of reproductive medicine, the first systematic review of the effectiveness of subfertility treatments was published in 1993 (Vandekerckhove et al., 1993).

In surgery a “non-evidence-based” approach to practice has been traditionally present (Johnson et al., 2008). The latest surgical technique is often embraced by the clinical community either when it seems rational or revolutionary or whenever it demonstrates the technical skill of the surgeon.

In this opinion paper we will present data and some conclusions on the current methodological quality of published reports of randomised studies in reproductive surgery. By discussing the benefits and pitfalls of RCTs in reproductive surgery, we aim to stimulate the reader’s interest in evidence-based practice in reproductive surgery.
Evidence-based medicine

“Evidence-based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients” (Sackett et al., 1996). The practice of evidence-based medicine (EBM) stands for the integration of individual clinical expertise with the best available external clinical evidence from systematic research. The essentials of EBM include five consecutive steps: first of all, to ask the right questions; secondly, to find the best level of evidence available; thirdly, to appraise critically the evidence for risk of bias, clinical relevance and applicability; fourthly, to implement the results of the appraisal in every day clinical practice and fifthly to evaluate the changes in practice (Farquhar and Vail, 2006). The highest level of evidence is derived from well written critically appraised systematic reviews of RCTs. The randomised controlled trial is generally accepted as being the least biased measure of the effectiveness of interventions. Although observational studies are considered vastly superior to RCTs in detecting adverse events e.g. surgical complications, they are often misleading when they are employed in searching for moderate treatment benefits. Systematic reviews comparing observational studies with randomised trials of the same interventions for the same conditions in the same study populations concluded that the former were clearly unreliable and consistently overestimated the treatment effect (Britton et al., 1998; Kunz et al., 2004).

RCTs in surgery: the benefits

Gynaecology has evolved to becoming a specialty in which the interventions are increasingly exposed to the gold standard of RCTs (Johnson et al., 2003). An overview of 23 systematic reviews including 94 gynaecological surgical trials in the Cochrane Database of Systematic Reviews (CDSR) (Selman et al., 2008) has ended up to the final conclusion that the quality of the RCTs has significantly improved since the Consolidated Standards of Reporting Trials (CONSORT) was introduced in 1996 (Begg et al., 1996). Using meta-regression analysis the authors have demonstrated that the proportion of studies reporting allocation concealment has significantly increased after the introduction of the CONSORT statement (60% versus 26%, p = 0.002). In parallel, a reduction in the magnitude of the effect estimate was observed over time (log of the ratio of odds ratios per year 0.96, 95% CI 0.93-0.99, p = 0.05) together with a trend towards higher precision of the estimation of the treatment effect (inverse of variance of the log odds ratio 0.12, 95% CI 0.02-0.23, p = 0.03) (Selman et al., 2008). In a second overview of 30 reviews in the CDSR, the same authors’ group found that only 7 out of 30 reviews reported evidence of a significant effect, 11 out of 30 reviews concluded that there was some evidence of significant effects for primary outcomes along with some evidence gaps while in the remaining 12, the authors found insufficient evidence of effectiveness (Johnson et al., 2008).

In conclusion, apart from providing up to date unbiased evidence on health care interventions, systematic reviews of RCTs can identify ‘gaps of knowledge’ where there is insufficient or no evidence at all. Several knowledge gaps in the evidence for fertility treatment have already been identified in a review of RCTs from the Cochrane Menstrual Disorders and Subfertility Group (MDSG) database (Johnson et al., 2003).

RCTs in surgery: the problems and pitfalls

There are two major categories of methodological challenges that need to be at least identified if not solved during the design phase of RCTs on surgical interventions (McCulloch et al., 2002; McLeod, 1999).

The first category concerns issues on the design and conduct of surgical trials. The surgical learning curve raises an interesting dilemma for the timing of surgical trials: it is well known that the individual surgeons’ complication rates fall significantly as the procedure is carried out on more and more patients. While drugs in trials work the same regardless of the competence of the prescribing physician, there are surgeon-to-surgeon differences in the preferences for and the expertise in performing different surgical procedures (Devereaux et al., 2005). In a recent Cochrane review on the effectiveness of excisional versus ablative surgery for ovarian endometrioma an effect favoring the excision of the cyst wall compared with its drainage and ablation was demonstrated the odds for a spontaneous pregnancy at 12 months after excision of the endometriotic cysts was higher compared with the control group which was treated by drainage and ablation (OR 5.2, 95% CI 1.9-14) (Hart et al., 2011). One of the included trials provided evidence for a treatment effect in favor of the excision technique for the spontaneous pregnancy rate at 12 months (OR 4.8, 95% CI 1.6-14, 62 patients) (Alborzì et al., 2004) while another smaller trial demonstrated a higher point estimate favoring excision over ablation but failed to reach statistical significance (OR 8.0, 95% CI 0.69-93, 26 patients) (Beretta et al., 1998). In the former trial, the intervention was performed by the same surgeon in two university centres, whereas in the latter no
information on the number of surgeons involved was available. Additionally, in both of them no information on the expertise of the performing surgeons was given. The need for head-to-head comparisons between different surgical techniques inevitably necessitates that the same surgeon prefers both techniques and is an expert in performing both of them. This is difficult in practice and impossible to achieve through studies. Therefore a strong case can be made for “expertise-based” trials in which consenting patients are allocated to different expert surgeons, who carry out the procedure they prefer and are expert in performing. While improving the internal validity of the trial, this potentially diminishes the external validity of the trial as well, meaning that the results of the RCT cannot be generalised as such without caution.

In the same context, in the same trials mentioned above, the application of both techniques raises further considerations: how sure can we be that both surgical teams were using comparable techniques? Did they selectively coagulate visible endometriotic lesions or was the whole cyst wall evaporated? No such clarifications were presented in the published reports. In addition, in one trial (Beretta et al., 1998) hydroflotation was used in contrast to the second trial (Alborzì et al., 2004). The previous remarks illustrate the great difficulty to standardize a surgical intervention since each individual surgeon develops his own modification of a standard technique e.g. for dissection, hemostasis and/or management of complications. There have been some attempts, though, to comprehensively standardise the technical steps of surgical interventions (Kapiteijn et al., 1999).

Another point which needs addressing is the difficulty of blinding or masking of a surgical procedure combined with the legal obligation of the treating physician to obtain informed consent. This can be a major problem if “soft” outcome measures, e.g. pain or quality of life are being assessed through self-reporting by unblinded patients or determined by unblinded assessors. The emotional consequences of knowing one’s treatment may significantly affect the reporting of outcomes. “Sham” surgical procedures have been conceived in the past to try to overcome the issue. Ethical problems may potentially arise (Moseley et al., 2002) and therefore “hard” outcome measures, e.g. live birth rate are mostly preferred. The latter are relatively independent to the knowledge of a patient’s treatment, but there still is the need for the outcome assessors to be blinded to the allocated treatments.

A second category of problems concerns the interpretation of the results of the trials. The mixture of data from trials conducted by less experienced surgeons together with others done by more expert ones may negatively affect the magnitude of the effect estimate, since differences in treatment outcomes are expected. In surgery, it is logical that a surgical intervention has a more favorable outcome when the provider is more experienced. Finally, a common problem in RCTs concerns the statistical power of surgical trials: a large survey of 90 “negative” surgical trials found that only 24% had sufficient power to detect relative risk reductions of 50% and only 29% reported a formal sample size calculation (Dimick JB et al., 2001). Power calculation is currently considered as an absolute must in the proper conduct of an RCT. It constitutes one of the main endpoints which a reviewer has to judge for a clinical trial and gives the adequate power to the results and therefore the interpretation of the trial’s data.

When is it ethical to design an RCT in surgery?

It is essential to define the circumstances under which an RCT can be conducted to determine whether a surgical procedure is more effective compared to other surgical or non-surgical treatments. We usually undertake trials because we hypothesise that a new surgical procedure can be better than the current standard practice in terms of efficacy, safety or cost but we are uncertain whether this statement is true or false. The limits of uncertainty include the possibility that the new technique may not be better or even worse than the current standard practice. The true uncertainty on the part of the expert professional community about the benefit to harms balance of two or more treatments for a well-defined study population has been described as the “clinical equipoise” (Freedman, 1987). When clinicians, methodologists and ethics committees or institutional review boards are uncertain whether an intervention is beneficial, an RCT is judged to be appropriate.

In addition we need to consider the uncertainty in the patient-clinician relationship, through an active patient participation in the inclusion/exclusion process. If the patient is certain that a specified treatment is better or safer, then the patient should not be included in the trial. Similarly, if the physician judges that a particular patient is clearly better off with a treatment, ethically, he is obliged to inform the patient and to assist in seeking the most appropriate treatment, excluding the patient from the trial. If both physician and patient are uncertain which treatment to choose, the patient should be offered inclusion in the trial.

The above principle of equipoise should always be considered as the gold standard in deciding whether or not to design an RCT. Some consider the evidence provided by non-randomised studies as an
ethical basis to discard the need for further research. Their certainty based on the results of studies with a high risk of bias should nevertheless be put aside in deference to the reasoned uncertainty existing within the larger community of experts (Haynes et al., 2006).

RCTs in reproductive surgery: the present state

Our group has published a systematic review on the effectiveness of reproductive surgery for treating female infertility (Bosteels et al., 2010). We conducted a search in the Cochrane Library, MEDLINE and EMBASE for RCTs on reproductive surgery in subfertile women. Our findings demonstrated a steady increase from 1970-2010 in the number of RCTs on the effectiveness of reproductive surgery per decade (Figure 1).

Nearly 75% of the included 63 RCTs had an adequate random sequence generation and nearly 50% had adequate allocation concealment (Figure 2). The percentage of RCTs on reproductive surgical interventions with adequate allocation concealment (26 out of 63 studies or 41%) was similar (p = 0.67) to the findings of the review of gynaecological surgical RCTs in the Cochrane Library (42 out of 94 studies or 45%) (Selman et al., 2008).

The number of trials with adequate random sequence generation has nearly doubled from the pre- to the post-CONSORT era (RR 1.7; 95% CI 0.98-3.1) (Figure 3); the difference was marginally insignificant (p = 0.06). Although the number of

Fig. 1. — RCTs on the effectiveness of reproductive surgery
to the treatment and the control group has been consistently shown to be the single most important factor in assessing the quality of RCTs (Farquhar and Vail, 2006). Nevertheless several large studies assessing the use of allocation concealment in different topic areas and subfertility trials have reported this item infrequently (Jüni et al., 1999; Moher et al., 1995; Schulz et al., 1994; Kjaergard et al., 2001).

This should be a major concern for trialists designing future RCTs in surgery. In contrast, while the absence of blinding is almost inherently associated with surgical trials, blinding has not been consistently shown to affect the estimation of the treatment effect magnitude (Jüni et al., 1999; Moher et al., 1995; Schulz et al., 1994; Kjaergard et al., 2001). The quality of the generation of the randomisation
sequence has similarly with the item of blinding not been shown to be of major importance in causing substantial bias (Jüni et al., 2001).

Considering the outcome measures, the majority of trials in subfertility and reproductive surgery do not report live birth outcomes as their primary outcome. This problem has already been highlighted by others (Vail and Gardner, 2003). It could be argued that all future trials on the effectiveness of reproductive surgical interventions should report live birth rates as their primary outcome since it is the single most important outcome of interest for couples undergoing fertility treatment. Ideally, the cumulative live birth rate, using life table analysis, should be described, as it accounts for the time to pregnancy and allows to subtract periods when the patient was not actively seeking to conceive. Time-to-event data are however troublesome for use in statistical pooling in meta-analyses. Moreover, the other outcome measures of interest in reproductive trials e.g. pregnancy and miscarriage rates should not be considered inferior since some conditions amenable to surgery may have an indirect impact on fertility, e.g. septate uterus which increases the probability of miscarriage.

Finally, the correct use of evidence statements should be encouraged. A common error observed in many studies is the confusion between “significant” and “important” or “clinically relevant”. A result is statistically significant if the difference observed between the study and the control samples is sufficiently convincing to signify a real difference in the population of which the sample is representative. A result is important or clinically relevant if the magnitude of the effect estimate is large enough to constitute a real difference between a control and study intervention for a given outcome. Ideally, authors and trialists should predefine minimally important clinical differences, based on estimates or trade-offs by physicians and/or patients of what really constitutes an important improvement of the outcome under study. If the sample size is large enough, a clinically unimportant or even trivial difference may signify a population difference, while in contrast clinically relevant differences may not be statistically significant if the sample size is too small. A second common error is the misinterpretation of a statistically non-significant finding. “Negative trials” do not exist! The correct expression of a conclusion is the absence of evidence of a particular effect and not the evidence of its absence (Altma and Bland, 1995; Alderson and Chalmers, 2003).

The methodological quality of surgical trials can be improved either through the training of surgeons in clinical epidemiology and evidence-based medicine or employing epidemiologists in surgical units where clinical research is being carried out (Urschel et al., 2001; Madhok et al., 2002). The evidence
from our recent systematic review is consistent with this viewpoint.

In conclusion, true progress in the field of reproductive surgery needs a balanced combination of surgical skills, a drive for innovation together with the exposure of clinical research to the undoubted validity of evidence-based medicine.

References


